Summative evaluation with a full-scale patient simulator: Challenges and adaptations

P.M. Sanderson,[1] M.O. Watson,[1] S. Jenkins,[2] D. Liu,[1] W.J. Russell,[2] N. Green,[3] P. Cole[1]
[1]ARC Key Centre for Human Factors, [2]Royal Adelaide Hospital, [3]Princess Alexandra Hospital

In this paper we outline considerations that went into designing and executing a full-scale simulator-based summative evaluation of four different display configurations for presenting information about anesthetized patients to an anesthesiologist. Although patient simulators appear to provide a "natural laboratory" for evaluating medical device innovations and equipment interface concepts, the software underlying patient simulators can be unequal to the challenges posed by the need for good representation of patient physiology and good experimental control. Moreover, the opportunities that full-scale patient simulators can offer for completely interactive, event-driven scenarios can present problems for experimental control and can promote participant hypervigilance. We describe the design of our experimental scenarios, the challenges our scenarios posed for simulator software and how we overcame those challenges, the design of a distractor task, and the methodology used to ensure we collected behavioral data sensitive to the manipulations of interest. Our adaptations in the face of challenges posed by the full-scale simulator context let us design an experiment that was highly informative about the advantages and disadvantages of the display configurations of interest.

Introduction

In addition to being used for education and training, full-scale patient simulators are increasingly used to evaluate how effectively equipment can be used by anesthesiologists (Drews, Syroid, Agutter, Strayer, & Westenskow, 2006). Human factors evaluations using patient simulators can help researchers detect latent errors and design faults early in the equipment development cycle or just prior to acquisition (Dalley, Robinson, Weller, & Caldwell, 2004). However, performing effective studies in a full-scale patient simulator has challenges, as we outline here.

In this study we wished to use a full-scale patient simulator to test the effectiveness of a respiratory sonification (continuous auditory display of an anesthetized patient's respiratory rate, tidal volume, and end-tidal carbon dioxide: see Watson & Sanderson, 2004, 2007). We also wished to test the relative effectiveness of the respiratory sonification (plus further auditory displays for blood pressure) against proximal visual displays, such as head mounted displays (HMDs). Prior evidence suggested that HMDs might provide similar advantages to advanced auditory displays (Sanderson, Watson, & Russell, 2005).

To compare the relative advantages of advanced auditory displays and HMDs we wished to compare anesthesiologists' ability to detect clinically significant patient events under the following four conditions in a relatively realistic anesthesia context.

- ***Visual***—Standard visual monitor with variable-tone pulse oximetry.
- ***HMD***—*Visual* plus HMD (monocular transparent Microvision™ Nomad)
- ***Audio***—*Visual* plus respiratory sonification (continuous two-tone auditory display of RR, ETCO2 and Vt) and blood pressure earcons (intermittent musical motifs for SBP and DBP from NIBP cuff)
- ***Both***—*Visual* plus *HMD* plus *Audio*.

In the following sections we outline challenges we faced and adaptations we had to perform in (1) the design of scenarios, (2) use of patient simulators, (3) control of participant attention, and (4) collection of behavioral evidence for event detection. Although each challenge by itself might appear simple to resolve, constraints of the full-scale simulator context meant that the combination of challenges was difficult to resolve.

The focus of the present paper is the above challenges, rather than the specifics of the actual displays evaluated. Similar challenges would arise for a wide range of equipment innovations.

Scenario challenges

To obtain adequate statistical power to compare the four display conditions noted above, we chose a within-subjects design. In order to conduct the within-subjects design and still complete each experimental run within half a day, we had to create a set of informative and well-balanced scenarios. Requirements for the scenarios were as follows.

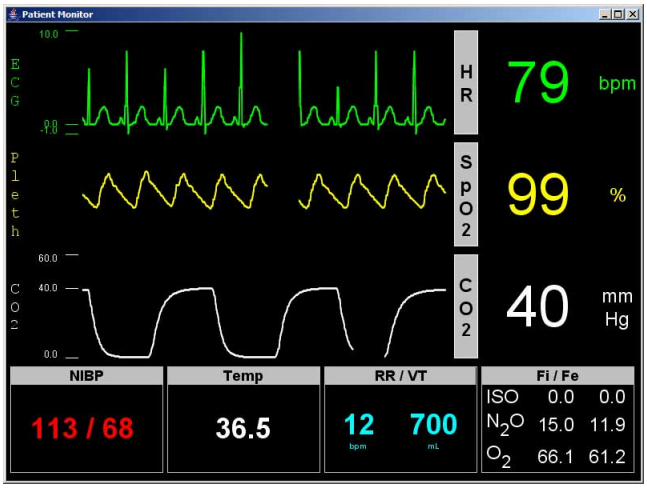

Figure 1: Visual monitor constructed for the simulator experiment. NIBP, plethysmography waveform (second from top), gas analysis, and capnography waveform (third from top) are evident.

First, each scenario had to run for the same amount of time (we eventually created 22-minute scenarios).

Second, the events in each scenario had to be detectable first by changes in respiratory vital signs such as respiratory rate, end-tidal carbon dioxide measures, or tidal volume and then (if blood pressure was involved physiologically) by blood pressure changes.

Third, each scenario had to contain three events so that three data points would be acquired per participant per display configuration, to increase the statistical power of the experiment. In addition, each scenario had to include one event whose resolution would not return the patient to the normal state.

Fourth, events had to be well separated in time for two reasons. First, we had to provide a time-window of up to about two minutes for the participant to report the event, especially as the time window needed to be long enough to capture event detections in conditions supporting slower detections. Second, there had to be a long period after the successful resolution of each event for the participant to be reassured that the participant had returned to normal and for normal monitoring to resume.

Fifth, if we allowed participating anesthesiologists to interact with the patient simulation software, then events might unfold in a manner incompatible with the above requirements. We therefore decided to create completely deterministic scenarios that would not require intervention by the anesthesiologist. We created a situation in which the anesthesiologist was required to supervise the activities of a junior colleague (eg an anesthesia resident). The anesthesiologist would be occupied with a distractor task that would have some degree of ecological validity (see Section on abstract classification task).

Scenarios were designed on paper that conformed to the above requirements. However it was not possible to program them with the METI ECS™ software because of the unavailability of control over certain vital signs and the tendency for the software not to reproduce exactly the same physiological sequences over multiple runs with the same input. We therefore made some simulator adaptations, as outlined in the next section.

Simulator adaptations

We observed major shortcomings of the simulator software (eg METI ECS™ and Body™) when used for display evaluations (Liu, Jenkins, Watson, Sanderson, & Russell, 2007). First, it can be difficult to manipulate patients in model-based simulators via indirect parameters such as drugs and fluids to achieve a specific and realistic pattern of physiological parameters (Cumin & Merry, 2007). Second, not all of the monitors used in the operating room are simulated. For example, Body™ does not provide a plethysmography waveform and the METI ECS™ lacks gas monitoring. Third, high-fidelity simulators provide more fidelity at the expense of less control or even no control over certain variables. Finally, many important aspects of monitor use are not simulated, such probe disconnection, leaks, and failures.

We developed software extensions to the BODY™ and METI ECS™ simulators and broadcast the results to our auditory and head-mounted display-based monitor prototypes over a TCP/IP-based protocol. Some of these extensions were used in the experiment reported here and the remainder were used in a subsequent experiment that is currently being analyzed. The software extensions let us incorporate the following five controller-driven simulated patient variables either during construction of scenarios or (in later experiments) while scenarios are being run.

- Non-invasive Blood Pressure (NIBP)
- Plethysmography
- Gas analysis
- Capnography
- ECG lead override.

In the present experiment, the METI ECS™ software and the Body™ software ran in parallel. The METI software controlled the patient manikin whereas the Body™ software controlled the display of vital signs on the patient monitors and display devices (eg Figure 1).

These capabilities made scenario design and development more flexible. They let us program a fuller range of events to compare the effectiveness of the advanced auditory monitoring and HMDs.

It is notable that because of the rich range of activities required, events witnessed, and activities observed, no participant expressed awareness during or after this

experiment that the scenarios had been deterministic and had not been driven by the activities and interventions of the junior colleague.

Attentional manipulations

As noted, to assure full experimental control, replicability across participants, and statistical power, our scenarios were completely deterministic. We wished to engage our participants' domain expertise and full involvement in the scenarios without letting them interact directly with the simulation software.

To achieve this we created a distractor task that would occupy anesthesiologists during the scenarios. Requiring participants to perform a distractor task had a further benefit. Anesthesiologists' vigilance levels are often higher in simulators than in the operating room because they expect adverse events to occur. Increased levels of vigilance could mean that results from simulator trials may not represent what would happen clinically when anesthesiologists are not hypervigilant.

Our abstract classification task controlled the level of distraction for anesthesiologists while they performed our scenarios. The abstract classification task (see Figure 2) required the anesthesiologist to read a series of scientific abstracts from *Anesthesia and Analgesia* and to classify each abstract according to the following:

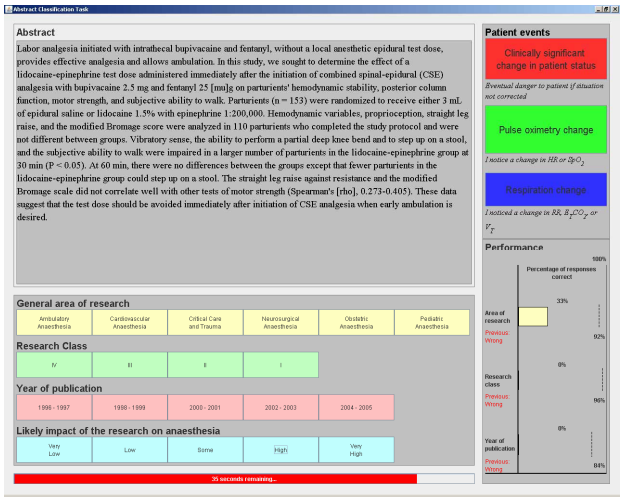

Figure 2: Abstract classification task. Abstract text is top left, four classification questions with answers to select from are bottom left. Buttons for registering patient events are top right (questions can vary given needs of experiments) and performance feedback for participant is bottom right.

(a) General area of research (6 options)
(b) Research evidence class (4 options)
(c) Year of publication (5 options)
(d) Likely impact of the research (5 options).

The arrival rate and conditions for turnover of each abstract can be varied. After some piloting, we set a forced pace of 40 seconds for the anesthesiologist to read and classify each abstract on all four categories. As a motivation to focus on the task, performance feedback was given on the screen after each abstract had been classified. As a further motivation, we provided "best performance to date" readouts that for the present experiment were set to be higher than pilot performance.

Performance of this task successfully simulated occasions when anesthesiologists become distracted in the operating room. Importantly, it sustained that level of distraction on a continuous basis to provide a controlled contrast of the effectiveness of the four display conditions in the face of anesthesiologist distraction.

Although the abstract classification distractor task is artificial it probably produces patient monitoring patterns that are closer to those in the operating theatre, where hypervigilance is less and many distractors are present, than to those in the simulator. It does so without the cost of requiring scores of hours of simulator time.

Figure 3: Participant, seated facing the distractor task, must turn to view visual patient monitor, to check activities of junior colleague (junior colleague is woman in top left and bottom left images), or to comment on an event to junior colleague. Red event detection button is visible in top right frame.

Data collection adaptations

The principal dependent variables for the experiment were whether the participants detected the adverse

patient events within the timeframe given, and how long they took to detect each event. A key concern was ensuring that event detections could be recorded as soon as possible after they occurred.

The most objective and standardised evidence could be collected by requiring participants to click the buttons provided on the distractor task screen (see top right of Figure 2—also in view in the top right frame of Figure 3). However there were many factors in the experiment that conspired to delay the emergence of behavioral evidence for event detections.

Because each abstract had to be classified within 40 seconds, the distractor task often held participants' attention until the classifications were completed, even though participants might have noticed a significant event. This created a delay in participants clicking the button to register that they had noted the event. The delay increased the likelihood that the event would be resolved before they had had a chance to register it.

Fortunately, participants often commented on the event to their junior colleague while they continued to work on the abstract classification task. The comment could be clearly picked up from the audiovisual recording (see Figure 3). Sometimes, however, participants would then forget to click the button on the distractor task screen. This meant that the earlier of either the button press or the verbal comment had to be taken as the detection time.

Figure 4: View from control room of scenario in progress. Participant is seated at left, partly hidden by drapes. Participant buddy is at centre rear near participant, writing on a clipboard. Anesthesia nurse is at right, partly obscuring participant buddy.

Participants would sometimes notice the event but would delay initiating communication with the junior colleague until the current abstract was fully classified, because the ensuing conversation might interrupt their ability to complete that abstract. If the participant also failed to press the significant event button, then there was no behavioral evidence for that detection.

As a result, the role of "participant buddy" was created to provide a third behavioral outlet that did not require the motor response with the mouse on the screen, and that did not require initiating communication with the junior colleague. An experimenter dressed in scrubs stood near the participant's workstation and the participant was told they could make comments, note events, express doubts and questions etc. to their "buddy", who would register any concerns (Figure 4).

During piloting the participant buddy role revealed a further reason that behavioral evidence for event detection might otherwise not eventuate. Sometimes participants would see or hear the start of an event, but would be unwilling to point it out to the junior colleague because they were not sure whether it was "significant" yet or whether the junior colleague might resolve it before the participant thought it was critical to point it out. These were response bias effects that varied across participants. Accordingly, participants were encouraged to make a brief verbal comment to the participant buddy if they had heard the start of an event but did not yet feel the need to point it out to the junior colleague.

As a result, the time of detection was taken to be the earliest of the following three behavioral indicators when an event was in progress:

- Button press on distractor task screen
- Comment to junior colleague about the event
- Comment to participant buddy about the event.

Substantive results

In our study we compared the four conditions outlined in the Introduction. Participants were 16 anesthesia registrars and consultants at Royal Adelaide Hospital. All participants served in four 22-minute anesthesia scenarios in a full-scale anesthesia simulator. Scenarios started with induction and proceeded through maintenance. All four display conditions were experienced by all participants and were presented in a counterbalanced order that varied across participants.

Participants performed the abstract classification task continuously while supervising the activities of their junior colleague. They were told that if they detected an anesthesia event that could harm the simulated patient, they either pressed the significant event button on the computer screen, informed their junior colleague

verbally, and/or informed their nearby "buddy" who recorded a response.

Compared with the Visual condition, participants detected significantly more events in the Audio and Both conditions (almost twice as many), but not in the HMD condition. Compared with the Visual condition, monitoring was rated as significantly easier in the HMD, Audio and Both conditions (Sanderson et al., 2007).

Participants generally showed equivalent levels of distraction across the four display conditions in terms of number of abstracts completed during each scenario (Watson et al., 2007). The fact that response accuracy was significantly better than chance suggests that they were paying attention to the task.

Responses to question (d) were not scored, as they simply involve forming an opinion. For questions (a) and (c), average performance was better than chance but did not change across displays. For question (b) about research evidence class, performance was again better than chance. Performance improved slightly as displays reduced the degree of continuous awareness participants had of the patient's state, so was best in the Visual condition where participants did not use any advanced displays and worst in the Both condition.

This suggests that performance on question (b) was resource-limited (Norman & Bobrow, 1975). It improved only when display configurations that were less continuously informative (such as the Visual display) "freed" cognitive resources from peripheral patient monitoring. This may indicate that the kind of patient monitoring supported in the Both condition and, to a lesser extent, in the Audio condition, was not as fully resource-free and "pre-attentive" as Woods suggested can happen for auditory displays (Woods, 1995). Further analysis will clarify this.

Conclusions

Thanks to the measures taken, our simulator-based experiment became a sensitive one statistically and an informative one in terms of our questions about different forms of advanced patient monitoring (Sanderson et al., 2007). However this outcome was achieved only through making adaptations to the way full-scale patient simulators are normally used, and by exploiting behavioral research tools and techniques to achieve the level of participant involvement required and capture representative and unbiased behavioral data.

It is clear that while they hold out the promise of letting researchers test device and display innovations, full-scale patient simulators can be quite limited due to the lack of control they offer over simulated patient status, the hypervigilance they create in anaesthesiologist participants, and the lack of ready measures of behavioral response to information sources. It is through experiments such as the one we have conducted that we start to define the requirements for full-scale patient simulators to be effective environments for behaviorally-oriented research into device innovations and interface design innovations.

Acknowledgements

This research was funded by an Australian Research Council Discovery Project DP0559504 to Sanderson, Watson, and Russell, by an APA postgraduate research award to David Liu, and by Royal Adelaide Hospital.

References

Cumin, D., & Merry, A. F. (2007). Simulators for use in anaesthesia. *Anaesthesia, 62*, 151-162.

Dalley, P., Robinson, B., Weller, J., & Caldwell, C. (2004). The use of high-fidelity human patient simulation and the introduction of new anesthesia delivery systems. *Anesthesia & Analgesia, 99*, 1737-1741.

Drews, F., Syroid, N., Agutter, J., Strayer, D., & Westenskow, D. R. (2006). Drug delivery as a control task: Improving performance in a common anesthetic task. *Human Factors, 48*(1), 85-94.

Liu, D., Jenkins, S., Watson, M., Sanderson, P., & Russell, W. J. (2007). *Extending simulators to improve support for patient monitoring display research (Abstract).* Paper presented at the Society for Technology in Anesthesia (STA2007), Orlando, FL, 17-20 January, 2007.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psych, 7*, 44-64.

Sanderson, P., Watson, M., & Russell, W. J. (2005). Advanced patient monitoring displays: Tools for continuous informing. *Anesth & Analg, 101*(1), 161-168.

Sanderson, P., Watson, M., Russell, W. J., Jenkins, S., Liu, D., Green, N., et al. (2007). *Advanced auditory displays and head-mounted displays: Advantages and disadvantages for monitoring by the distracted anesthesiologist (Abstract).* Paper presented at the Society for Technology in Anesthesia (STA2007), Orlando, FL, 17-20 January, 2007.

Watson, M., & Sanderson, P. (2004). Sonification helps eyes-free respiratory monitoring and task timesharing. *Human Factors, 46*(3), 497-517.

Watson, M., & Sanderson, P. (2007). Designing for attention with sound: Challenges and extensions to Ecological Interface Design. *Human Factors, 49*(2).

Watson, M., Sanderson, P., Russell, W. J., Llewelyn, K., Liu, D., & Lacherez, P. (2007). *Simulating high workload situations to evaluate patient monitors (Abstract).* Paper presented at the Society for Technology in Anesthesia, Orlando, FL, 17-20 January, 2007.

Woods, D. D. (1995). The alarm problem and directed attention in dynamic fault management. *Ergonomics, 38*(11), 2371-2393.